

Project Overview

This project aims to study and predict the differences between demand and supply for a transportation service in different areas over a certain time frame. The goal is to evaluate how well various prediction models can estimate these differences based on past data.

Data Preparation

The dataset comprises order information with features such as region identifiers, day numbers, time slots, demand, and supply values. The dataset spans 21 days, and our goal is to use this historical data to forecast the gap for the next 7 days. The data preparation steps include:

Loading Data: Importing order data and region mapping data into a structured format using pandas.

Merging Data: Combining order data with region information to align each order with its corresponding geographical area.

Feature Engineering: Extracting useful features such as day number and time slot from the timestamp data to prepare the dataset for modeling.

Model Training and Evaluation

Several regression models were employed to predict the demand-supply gap:

Linear Regression: Used as a baseline model.

K-Nearest Neighbors (KNN): This model doesn't rely on assumptions about data patterns and is useful for understanding complex data relationships.

Random Forest: This approach combines multiple models and is highly accurate, with a strong ability to avoid fitting the data too closely.

Lasso Regression: Utilized to observe the impact of regularization on feature selection.

Ridge Regression: Employed to compare its regularization effects with Lasso.

Each model was trained on the dataset, and predictions were evaluated using the R-squared value to determine how well the model explains the variability of the response data, and RMSE (Root Mean Square Error) to measure the average magnitude of the prediction errors.

	region_id	day_number	time_slot	demand	supply	gap
0	1	1	1	187	178	9
1	1	1	2	198	191	7
2	1	1	3	192	182	10
3	1	1	4	172	167	5
4	1	1	5	153	152	1

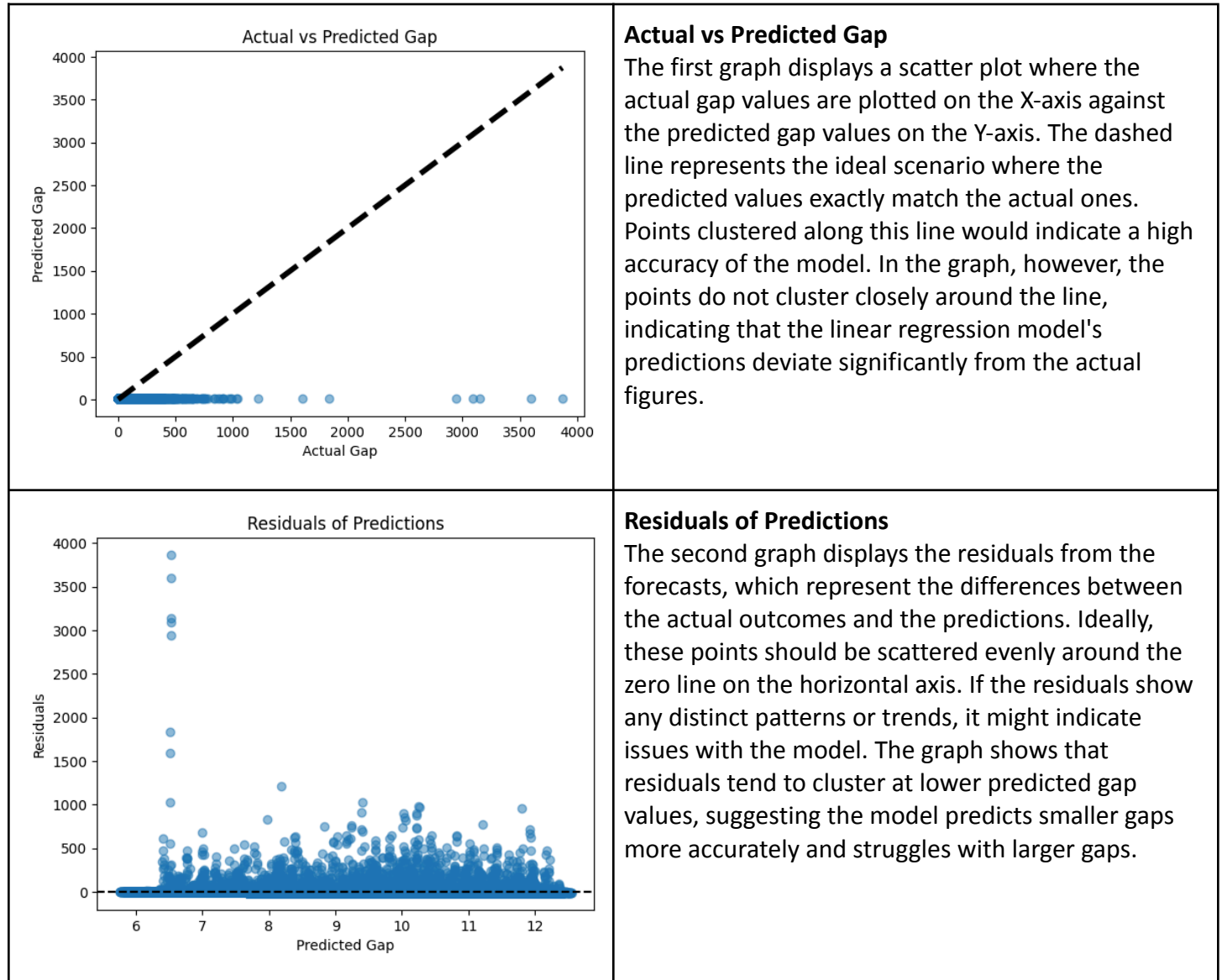
Results

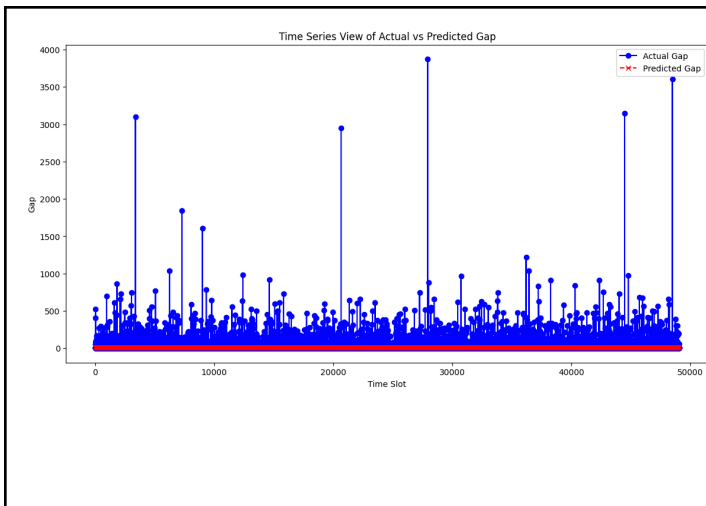
Linear Regression

R-squared: 0.0005478037494272003

RMSE: 54.56

The Linear Regression model shows an extremely low R-squared value, indicating that it is almost ineffective in explaining the variance in the demand-supply gap across the dataset. The high RMSE value further suggests that the predictions are, on average, 54.56 units away from the actual data points. This performance indicates that Linear Regression, perhaps due to its inability to handle non-linear relationships and interactions between features effectively, is not suitable for this particular dataset.





Time Series View of Actual vs Predicted Gap

The third graph presents a time series plot with the actual gaps shown in blue and the predicted gaps in red. In an ideal model, both lines would overlap considerably. The graph shows that the predicted gaps often fall below the actual gaps, and there is substantial variability in the accuracy of the predictions across different time periods. Some peaks are predicted reasonably well, while others are missed, indicating inconsistency in the model's performance over time.

Findings

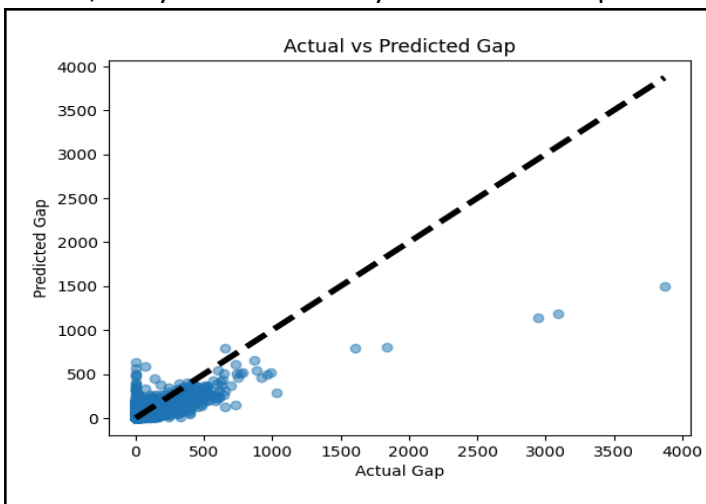
Overall, the findings from these graphs suggest that while the linear regression model can provide a rough estimate of the demand-supply gap, its predictions are not consistently reliable, especially for higher values of the gap. The variability in prediction accuracy across different time slots also indicates that the model does not capture all the factors influencing the demand-supply dynamics or that these factors may vary significantly over time. The insights from these charts show that we need a more advanced model or extra features to enhance the accuracy of our predictions.

K-Nearest Neighbors (KNN)

R-squared: 0.5993399492221101

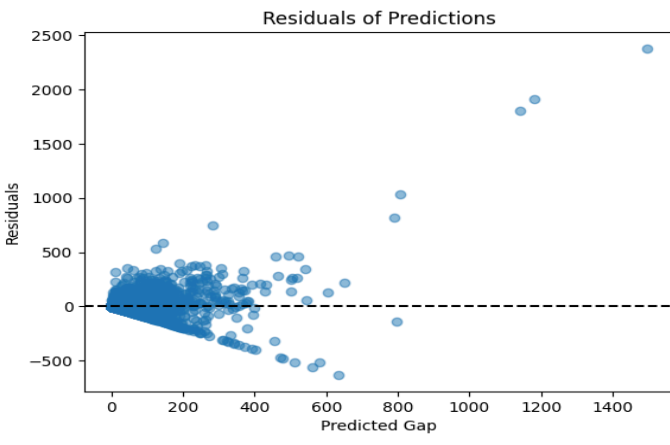
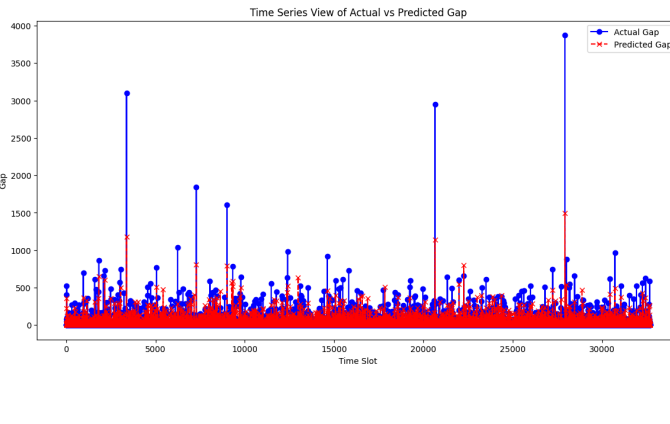
RMSE: 33.52

The KNN model shows a substantial improvement over Linear Regression. With an R-squared value of 0.5993, it explains approximately 60% of the variance in the gap data, which is a significant increase. The RMSE is also considerably lower, indicating that the predictions are closer to the actual data points. The performance of KNN suggests that the model is capturing more complex relationships in the data that Linear Regression misses, likely due to its ability to consider the proximity of data points in its predictions.



Actual vs Predicted Gap for KNN

This graph is a scatter plot that shows each actual gap value from the dataset on the X-axis against the corresponding predicted gap value on the Y-axis. The dashed diagonal line represents the line of perfect prediction. The concentration of points along this line would indicate accurate predictions. Here, we observe a cloud of points that shows a trend toward this line, suggesting that the KNN model has a moderate level of predictive accuracy, although there are still considerable deviations, particularly with larger gap values.

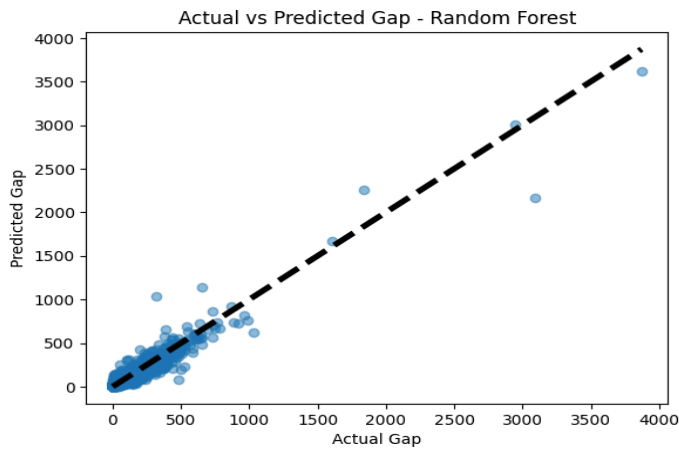
 <p>A scatter plot titled 'Residuals of Predictions' showing the residuals of KNN predictions against the predicted gap values. The x-axis is labeled 'Predicted Gap' and ranges from 0 to 1400. The y-axis is labeled 'Residuals' and ranges from -500 to 2500. A dashed horizontal line at y=0 represents zero residual. Most data points are clustered near the zero line for predicted gaps below 600, but for predicted gaps above 800, the residuals are significantly positive, reaching up to 2000, indicating a systematic overestimation for larger gaps.</p>	<p>Residuals of Predictions for KNN</p> <p>The second graph depicts the residuals or errors between the predicted and actual values, plotted against the predicted gap values. Ideally, we would expect to see the points randomly dispersed around the horizontal line at zero, indicating that the model's errors are not systematic. The plotted points show a cluster near the lower end of the predicted gap range, which indicates that the model predicts smaller gaps with more accuracy than larger gaps.</p>
 <p>A time series plot titled 'Time Series View of Actual vs Predicted Gap' showing the actual gap (blue line with diamond markers) and the predicted gap (red line with 'x' markers) over 30,000 time slots. The y-axis is labeled 'Gap' and ranges from 0 to 4000. The predicted gap line follows the general trend of the actual gap line but consistently underestimates the magnitude of the spikes, particularly for gaps exceeding 1000.</p>	<p>Time Series View of Actual vs Predicted Gap for KNN</p> <p>The third graph is a time series plot with the actual gaps plotted against the predicted gaps over sequential time slots. The actual gaps are marked in blue, and the predicted gaps are marked in red with a line connecting them, indicating the time sequence. The plot reveals that while the KNN model captures the trend and some spikes in the demand-supply gap reasonably well, it tends to underestimate the size of the larger gaps.</p>
<p>Findings:</p> <p>These graphs suggest that the KNN model, with its ability to consider the closest neighboring points in the dataset, provides a more nuanced prediction of the demand-supply gap compared to linear regression. It seems better at capturing certain patterns in the data, especially for smaller gaps. However, the model's difficulty in accurately predicting larger gaps is evident and may be due to outliers or the effects of variables that are not well-represented by the nearest neighbors approach. The performance depicted by these visualizations indicates that while KNN is a more effective model than linear regression for this task, there is still room for improvement, potentially through parameter tuning or incorporating additional features into the model.</p>	

Random Forest

R-squared: 0.9278693160272794

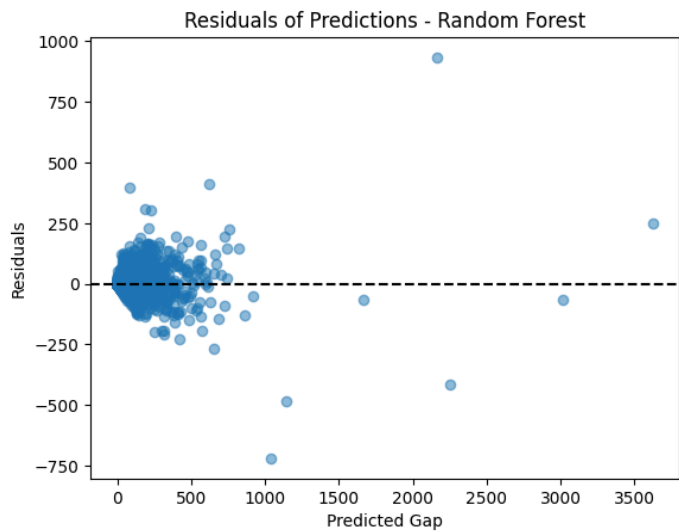
RMSE: 14.22

Out of all the models tested, Random Forest performs the best. With an R-squared score of 0.9279, it is highly predictive and appropriate for this set of data, explaining almost 92% of the variance. With a tiny average deviation, the low RMSE value suggests that the predictions are fairly close to the actual values. The ensemble approach used in this model, which combines several decision trees to create a more robust and generalized model that can handle outliers and other types of data anomalies, is responsible for its success.



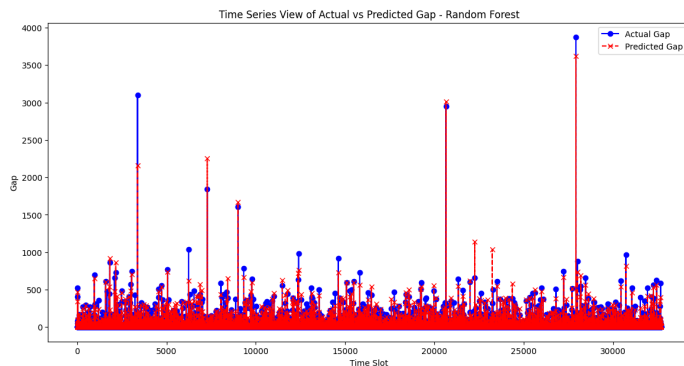
Actual vs Predicted Gap - Random Forest

This graph plots actual gap values against the predicted ones. The dashed diagonal line represents the line of perfect predictions where the actual values match the predicted ones. The points are notably closer to the line compared to previous models, indicating a much higher accuracy in the Random Forest's predictions. However, there are still some instances, particularly at higher gap values, where the model's predictions deviate from the actual values.



Residuals of Predictions - Random Forest

The second graph shows the residuals, which are the differences between the actual values and the model's predictions. The residuals appear to be more tightly clustered around the zero line and display a slight fan-shaped dispersion as the predicted gap increases. This suggests that the model is highly accurate for smaller gaps but less so for larger ones. The presence of a few large residuals indicates some outliers or instances where the model's predictions are significantly different from the actual gaps.



Time Series View of Actual vs Predicted Gap - Random Forest

The final graph is a time series representation of actual and predicted gaps. The blue line represents the actual gap, and the red line with markers shows the predicted gap. This graph highlights the Random Forest's effectiveness in tracking the fluctuations in the gap over time. It successfully captures the overall trend and the occurrence of peaks, albeit with some discrepancies in the peak magnitudes.

Findings:

The closeness of data points to the line of perfect prediction and the clustering of residuals around zero indicate the excellent predictive power of the Random Forest model. The demand-supply gap's overall trend and sporadic spikes may be reasonably predicted by the model, which is well-tuned to handle the data's complexity. Nonetheless, the performance on larger gaps and the presence of outliers suggest that there could be room for further improvement, possibly by feature engineering or hyperparameter optimization.

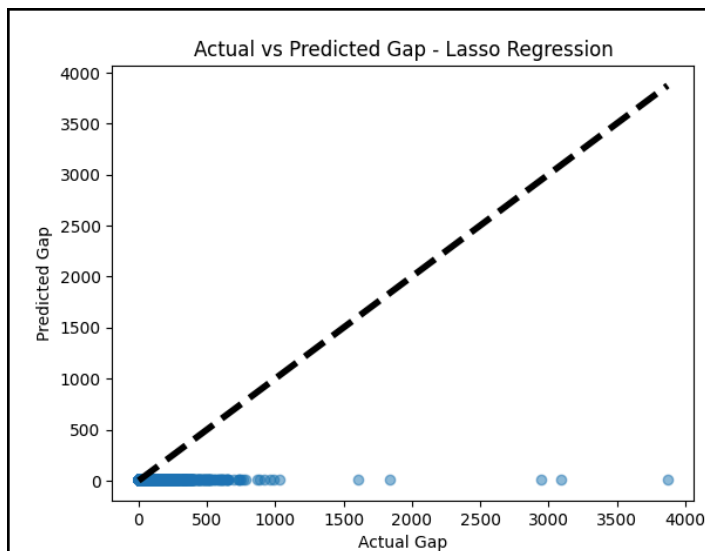
The time series graph confirms the model's capacity to grasp temporal patterns, reinforcing its suitability for forecasting in dynamic settings.

Lasso Regression

R-squared: 0.000786277562049964

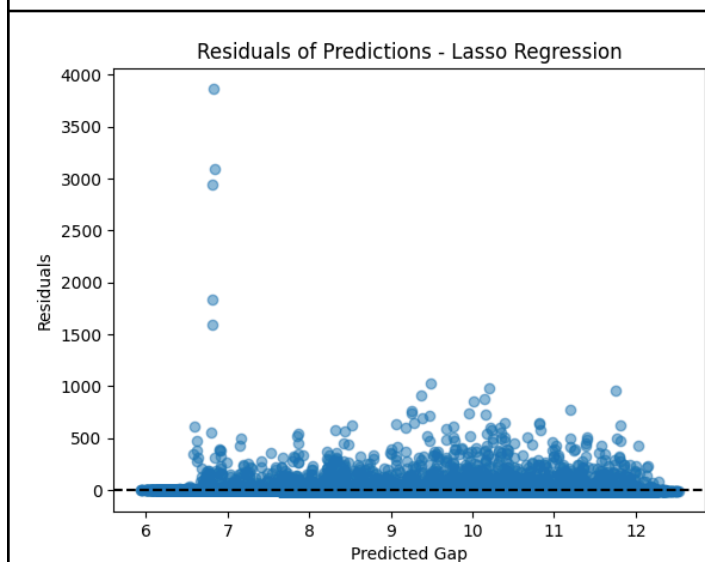
RMSE: 52.93

Lasso Regression performs slightly better than Linear Regression in terms of R-squared but still remains practically ineffective for this application. The small improvement in R-squared value suggests that some minimal feature selection provided by Lasso's regularization did not significantly impact the model's ability to predict the gap. The RMSE remains high, almost similar to that of Linear Regression, indicating that the predictions are still not close to the actual values.



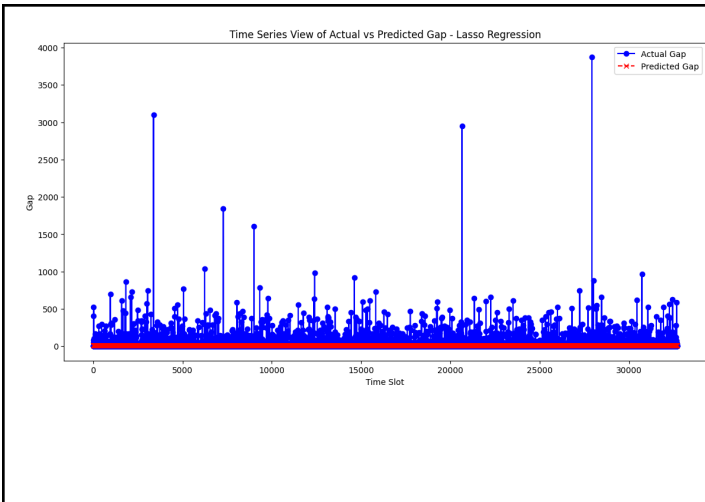
Actual vs Predicted Gap - Lasso Regression

The first graph is a scatter plot of the actual gaps against the predicted gaps. The dashed line represents the line where the predicted values are equal to the actual values. The data points are concentrated at the lower end of the gap values and spread out as the gap values increase, showing that the Lasso model's predictions are more accurate for smaller gaps. The model does not predict the larger gaps well, as evidenced by the data points that diverge from the diagonal line.



Residuals of Predictions - Lasso Regression

The second graph displays the residuals of the Lasso model's predictions, which are the differences between the actual and predicted gap values. Ideally, we would expect to see a random spread of residuals around the horizontal line at zero. Instead, we observe a concentration of residuals for lower predicted values and a spread for higher ones, indicating that the prediction error increases with the size of the gap.



Time Series View of Actual vs Predicted Gap - Lasso Regression

The third graph shows a time series of actual and predicted gaps over different time slots. The actual gaps are marked with blue lines, and the predicted gaps with red markers. It appears that the Lasso model, which includes a penalty to reduce overfitting by shrinking coefficients, has difficulty capturing the larger fluctuations in the data, leading to underestimation of the larger peaks in the demand-supply gap.

Findings:

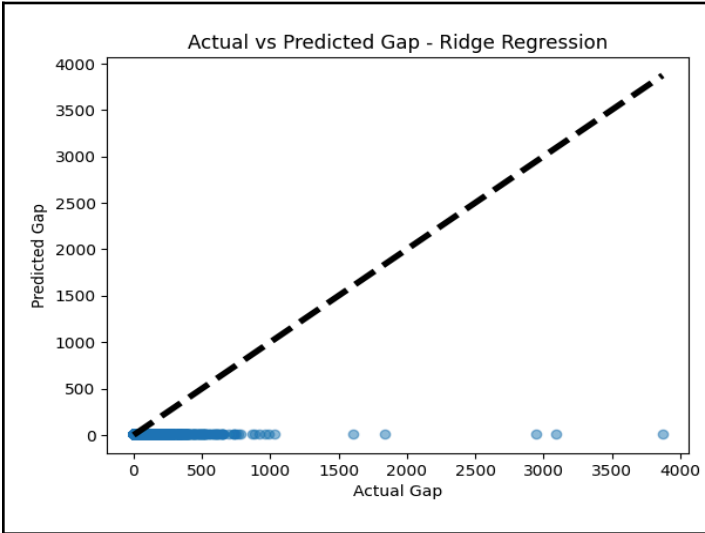
These findings show that, although the Lasso regression model may recognize a certain degree of the underlying patterns in the supply-demand gap, its overall forecasting ability is constrained. The shrinkage strategy used by the model may be the cause of its inability to predict greater gaps properly. This method has the potential to overly penalize and lower the coefficients of crucial predictors. When a model is overly straightforward to accurately represent the underlying intricacies of the data, it might lead to underfitting. All things considered, the graphics point to the possibility that the Lasso model is not the best option for this dataset and the current prediction task—especially if the objective is to predict larger fluctuations in the demand-supply gap.

Ridge Regression

R-squared: 0.0007871914482450171

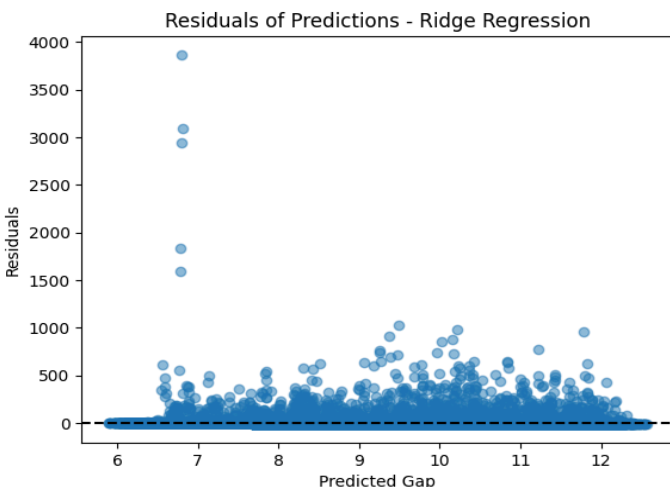
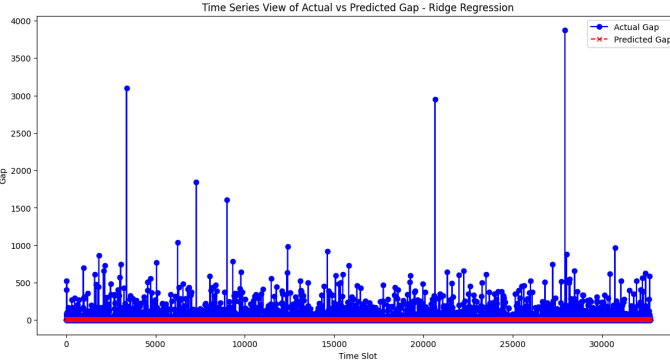
RMSE: 52.93

Ridge Regression, like Lasso, shows only a negligible improvement over Linear Regression. The similarity in performance between Lasso and Ridge in this context indicates that regularization alone (without other model adjustments) is insufficient to capture the complexity of the dataset. The RMSE being similar to that of Linear and Lasso Regression reinforces the conclusion that Ridge Regression is not effective for this dataset.



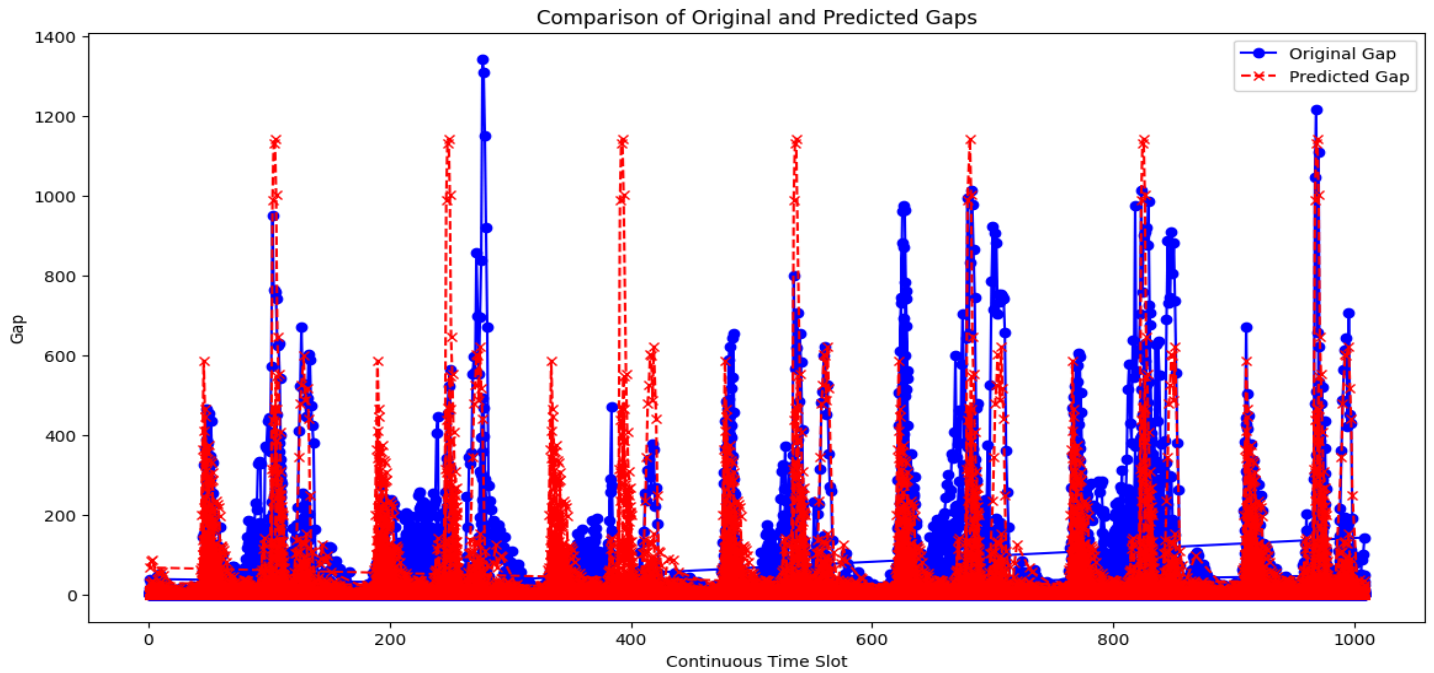
Actual vs Predicted Gap - Ridge Regression

This graph shows the actual gaps on the X-axis against the predicted gaps on the Y-axis, with a dashed line representing the line of perfect prediction. The data points are closer to the origin and don't extend far along the line, which suggests that Ridge Regression predicts smaller gaps with reasonable accuracy. However, for larger gaps, the predictions deviate significantly, indicated by the scarcity of points near the line as the gap values increase.

 <p>A scatter plot titled 'Residuals of Predictions - Ridge Regression'. The y-axis is labeled 'Residuals' and ranges from 0 to 4000. The x-axis is labeled 'Predicted Gap' and ranges from 6 to 12. The plot shows a dense cluster of blue points near the x-axis (residuals near 0) for predicted gap values between 6 and 8. As the predicted gap increases beyond 8, the points spread out vertically, with several notable outliers reaching residuals up to 4000 at a predicted gap of approximately 7.</p>	<p>Residuals of Predictions - Ridge Regression</p> <p>The second graph illustrates the residuals, the difference between the actual gaps and the predicted gaps, plotted against the predicted values. The residuals are mostly clustered near the start of the axis, indicating lower prediction errors for smaller gaps. As the predicted gap increases, the residuals spread out, showing that errors become more pronounced, particularly for larger gap predictions.</p>
 <p>A time series plot titled 'Time Series View of Actual vs Predicted Gap - Ridge Regression'. The y-axis is labeled 'Gap' and ranges from 0 to 4000. The x-axis is labeled 'Time Slot' and ranges from 0 to 30000. The plot shows two data series: 'Actual Gap' represented by blue vertical spikes and 'Predicted Gap' represented by a red line with 'x' markers. The predicted gap line remains consistently low, mostly below 500, while the actual gap spikes frequently, with several major peaks reaching values between 1500 and 3500.</p>	<p>Time Series View of Actual vs Predicted Gap - Ridge Regression</p> <p>The final graph is a time series plot, marking the actual gap values over time with blue lines and the predicted gaps with red markers connected by lines. While the Ridge Regression model appears to track the frequency of the gaps, it tends to underestimate their magnitude, especially the peaks, which suggests that the model might not capture the more extreme variations in the demand-supply gap.</p>
<p>Findings:</p> <p>These visuals suggest that Ridge Regression, which incorporates regularization to reduce overfitting by penalizing large coefficients, provides a modest prediction of the demand-supply gap. It performs better at forecasting smaller gaps but is less accurate with larger ones. This could be due to the model's inherent bias towards smaller coefficient values, which can hinder its ability to capture more significant trends and peaks in the data. The time series graph reinforces the model's challenge in accurately predicting times of high demand-supply fluctuation. While Ridge Regression offers some predictive insights, it may require further tuning or the integration of more complex features to improve its performance for forecasting purposes.</p>	

Forecasting and Visualization

For forecasting, the Random Forest model was used due to its superior performance. We replicated last week's data to predict the future week's demand-supply gap, using historical demand and supply patterns as a basis. The forecasted data was visualized to compare the predicted gaps against actual data, providing a clear depiction of model effectiveness in a real-world scenario.



Graph: Comparison of Original and Predicted Gaps in Forecasting

The blue markers denote the actual gaps, while the red markers connected by dashed lines depict the predicted gaps. This visual comparison is critical for assessing the model's ability to forecast future values based on past data.

The distribution of blue and red markers across the time series suggests a rhythmic pattern of highs and lows, reflective of demand-supply dynamics over time. Peaks in the actual data are mirrored by the predictions, indicating the model's general ability to detect when larger gaps are likely to occur.

Despite the model capturing the pattern's rhythm, there are clear differences in peak magnitudes. In several instances, the predicted gaps underestimate or overestimate the actual gaps. This discrepancy could be due to the model's limited capacity to account for all influencing factors or an inherent time lag in its predictive responses.

Conclusion

The Random Forest model outperformed other models significantly, making it the most suitable choice for this predictive task. Its ability to handle various data types and complex relationships between features effectively captured the underlying patterns in the demand-supply data.